Written Comprehensive Examination

Methods 210, 211, 212

Department of Statistics, UC Irvine Monday, June 22, 2020, 9:00 am to 12:00 pm

- There are 4 questions on the examination. You are to do 3 of 4 questions.
- Your solutions to each problem should be written on separate sheets of paper. Label each sheet with your student identification number, the problem number, and the page number of that solution written in the upper right hand corner. For example, the labeling on a page may be:

ID# 912346378
Problem 2, page 3

- You have 3 hours to complete your solution. Please be prepared to turn in your exam at 12:00pm.

1. Salary inequities by gender continue to be a problem in academia, government, and industry. In this problem we will consider an analysis of monthly salaries for faculty from a single R1 university in the US during a single year. In total, monthly salary data, denoted Y. was obtained on $n = 1,597$ faculty members. The primary goal of the analysis is to determine whether or not evidence for gender discrimination exists with respect to pay. Along with the monthly salary values, additional covariates will be introduced into the question throughout.

   (a) We will start by considering a regression model of the following form:

   $$Y_i = \beta_0 + \beta_1 I_{\texttt{male}_i=1} + \epsilon_i, \quad i = 1, \ldots, 1597.$$

   In matrix notation, this can be written as

   $$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon},$$

   with $\vec{Y}$ denoting the $n \times 1$ vector of monthly salaries, $\mathbf{X}_{n\times 2} = (\vec{1} \quad \vec{I}_{\texttt{male}_i=1})$, $\vec{\beta} = (\beta_0, \beta_1)$ a 2×1 column vector, and $\vec{\epsilon}$ an $n \times 1$ column vector of model residuals. Using the matrix formulation of the problem, derive the ordinary least squares estimator of $\vec{\beta}$, which we will denote by $\widehat{\vec{\beta}}$. Leaving your solution for $\widehat{\vec{\beta}}$ in matrix notation is perfectly fine.

   (b) The classical linear regression model assumes $\vec{\epsilon} \sim N_n(\vec{0}, \sigma^2 \mathbf{I}_{n\times n})$. Under this assumption, derive the variance of $\widehat{\vec{\beta}}$.

   (c) Below is R output from fitting the model in (a) via OLS to the available salary data. Provide a precise interpretation of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (or some suitable transformation of these parameters) in words that can be understood by a statistical layman.

```
##
##### OLS fit
##
> fit1 <- lm( salary ~ i.male, data=salary )
> summary(fit1)

Call:
lm(formula = salary ~ i.male, data = salary)

Residuals:
   Min     1Q Median     3Q    Max
 -3601  -1406   -419   1091   7732

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5396.9       96.5    55.9   <2e-16 ***
i.male       1334.7      111.9    11.9   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1950 on 1595 degrees of freedom
Multiple R-squared:  0.0819,Adjusted R-squared:  0.0813
F-statistic:  142 on 1 and 1595 DF,  p-value: <2e-16
```

   (d) Using the output from (c), provide a 95% confidence interval for β_1.

(e) Based upon these output, a 95% Wald-based confidence interval for the mean monthly salary among male faculty members is (6620.5, 6842.7). Use this and the model output in (c) to obtain an estimate of the covariance between $\widehat{\beta}_0$ and $\widehat{\beta}_1$. (You may leave your expression unevaluated, but your estimate should be a function of the output given.)

(f) Suppose that in truth $Y_i \sim Exp(\mu_i)$ with $\mu_i = \beta_0 + \beta_1 I_{\texttt{male}_i\texttt{=1}}$ but we still compute the OLS estimator for the model

$$\mathrm{E}[\vec{Y}] = \vec{\mu} = \mathbf{X}\vec{\beta},$$

where $\vec{Y}$, $\mathbf{X}$ and $\vec{\beta}$ are as defined in (a). Again using matrix notation, derive the mean and variance of the OLS estimator in this case. From these results, state which estimates from the model fit in (c) can be "trusted" and which cannot.

(g) Again consider the setting in (f). Based upon the model output in (c), state which of the following are true (in large samples like we have) and which are false. Briefly provide the reason for your response in each case.

(1) A test of $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ will be valid in that it will yield approximately the nominal type I error rate.

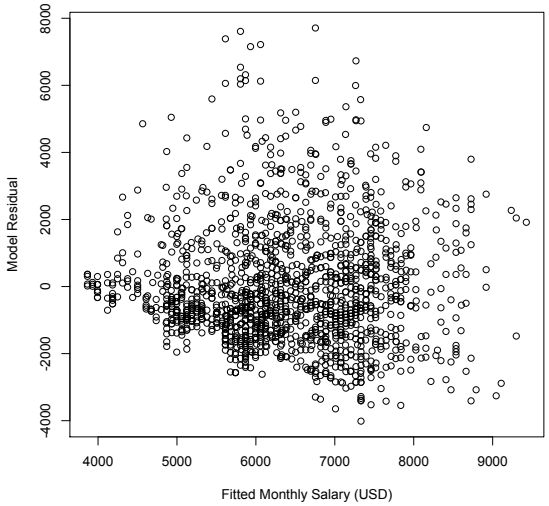
(2) A 95% confidence interval for β_1 will be valid in that it will yield approximately correct coverage probability.

(h) A colleague noted that the model in (a) is not likely to provide a fair comparison of the mean salaries between males and females with the goal of assessing whether the university is guilty of systematically paying males more the females. She suggests that, at minimum, your model should include the years since the faculty member was hired at the university and the primary school for their appointment (categorized as 1=Humanities/Social Sciences, 2=Engineering/CS, 3=Physical Sciences/Biological Sciences, 4=Business/Law/Medicine). Briefly explain why your colleague makes a valid point.

(i) Suppose you modify the model in (a) based upon the above suggestions and again use OLS to fit the following:

$$Y_i = \gamma_0 + \gamma_1 I_{\texttt{male}_i\texttt{=1}} + \gamma_2 \texttt{yr.since.hired} + \gamma_3 I_{\texttt{school}_i\texttt{=2}} + \gamma_4 I_{\texttt{school}_i\texttt{=3}} + \gamma_5 I_{\texttt{school}_i\texttt{=4}} + \epsilon_i.$$

Below is a plot of the residuals from the OLS fit of this model vs. the fitted salary values from the model. Based upon this, does it appear that the assumptions of classical linear regression are valid for these data? Explain. If not, provide two ways the model can be *changed* to address any of the violations you believe are reflected in the plot.



(j) After performing residuals diagnostics you find that the distribution of log-transformed salary is symmetric and modify the model in part (i) as follows:

$$\ln(Y_i) = \delta_0 + \delta_1 I_{\texttt{male}_i\texttt{=1}} + \delta_2 \log_2(\texttt{yr.since.hired}) + \delta_3 I_{\texttt{school}_i\texttt{=2}} + \delta_4 I_{\texttt{school}_i\texttt{=3}} + \delta_5 I_{\texttt{school}_i\texttt{=4}} + \epsilon_i.$$

(1) Provide a precise interpretation of δ_1 (or a suitable transformation of this parameter) in words that can be understood by non-technical individuals. (Note that your interpretation should likely not involve $\log_e$-salary in order to be understandable!)

(2) Explain how the interpretation of δ_1 and the interpretation of γ_1 (from part (i)) differ and what the relevance of this difference is in terms of your question of interest.

(3) Provide a precise interpretation of δ_2 (or a suitable transformation of this parameter) in words that can be understood by non-technical individuals. (Note that your interpretation should likely not involve $\log_e$-salary or $\log_2$-years since hired in or to be understandable!)

END OF QUESTION (1)

2. A microbiology lab is studying effects of two treatments of bacterial skin infection. We will refer to them as treatments A and B.

   (a) First, lab researchers designed and performed the following experiment. They cultured bacterial strains of interest in petri dishes until each culture reached the mid-exponential growth phase, with 30 of these cultures receiving no treatment (control group), 30 receiving treatment A, and 30 receiving treatment B. Assignment to treatments and control groups was done uniformly at random. After an over-night incubation period, researchers recorded the number of colony forming units (CFUs) measured in millions per mL in each petri dish. They assume that all petri dishes had the same number of CFUs at the mid-exponential growth phase. If a treatment works, the researchers expect to see lower number of CFUs after the over-night incubation in the treated petri dishes than in the untreated/control ones. Taking a regression view of this problem, let Y_i be post-treatment or control number of CFUs for petri dish i, a_i be a binary indicator of treatment A (1 = treatment A applied, 0 = otherwise), b_i be a binary indicator of treatment B (1 = treatment B applied, 0 = otherwise), c_i be a binary indicator of control group (1 = no treatment applied, 0 = otherwise).

      i. We assume the following linear model:

$$Y_i = \beta_0 + \beta_1 a_i + \beta_2 b_i + \epsilon_i, i = 1, \ldots, 90,$$

      where $\epsilon_i \sim N(0, \sigma^2)$. We assume that normality assumption holds here. Using ordinary least squares estimation (OLS), we obtain $\hat{\beta}_0$ = 2.2, $\hat{\beta}_1$ = -0.21, $\hat{\beta}_2$ = -0.13. Provide interpretation of these coefficients.

      ii. For the above regression, $\widehat{MSE} = 0.07$ and

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.03 & -0.03 & -0.03 \\ -0.03 & 0.07 & 0.03 \\ -0.03 & 0.03 & 0.07 \end{pmatrix}.$$

      Compute 95% confidence intervals for β_1 and β_2, using the fact that $\Pr(Z > 1.99) \approx 0.975$, where $Z \sim t_{87}$. State your conclusions about anti-bacterial effects of treatments A and B.

      iii. Compute 95% confidence interval for $\beta_1 - \beta_2$ and interpret it.

      iv. Provide a different linear model formulation that would allow you to compute 95% confidence interval for $\beta_1 - \beta_2$ and to test H_0: $\beta_1 - \beta_2 = 0$ directly from the OLS output, without the need to know $(\mathbf{X}^T\mathbf{X})^{-1}$.

   (b) Since treatments A and B use different biological mechanisms to inhibit bacterial growth, the researchers became curious about the effect of applying both treatments simultaneously, hoping to find treatment synergy — treatments having a larger effect when applied simultaneously than the sum of their individual effects. They repeated the above experiment, but added a forth group of petri dishes that recieved both treatments A and B.

      i. Formulate a linear model that will allow the researchers to estimate an effect of synergy or competition (opposite of synergy). Provide interpretation for all model coefficients.

ii. Explain what test you would apply to check for existence of a synergistic effect between the two treatments.

(c) When the researchers looked at the data more carefully, they noticed that they were not able to "catch" all petri dishes exactly at the mid-exponential growth phase, so the starting numbers of CFUs for all petri dishes were not identical. Let's denote this starting number of CFUs for petri dish i by x_i.

i. Formulate an analysis of covariance (ANCOVA) model that would allow the researchers to estimate effects of treatments A and B (without synergy considerations), while accounting for different initial conditions (starting numbers of CFUs). What parameter estimates, confidence intervals, and hypothesis tests would you report to the researchers?

ii. Compare and contrast assumptions of the ANCOVA model above and an ANOVA model with response being $Y_i - x_i$.

iii. Describe how you would perform model diagnostics for your ANCOVA model.

iv. Suppose your diagnostic analysis reveals problems with the ANCOVA model. You come with these results to the microbiologists and they tell you that the root of the problem could be in the anti-bacterial treatment mechanisms, which are not fully understood. The treatments can either kill some unknown number of bacteria or diminish the *rate* at which bacteria grow. Extend your ANCOVA model to allow for both anti-bacterial mechanisms and explain how you would use this new model to help the researchers decide which mechanism each treatment uses.

v. Comment on what could go wrong with the ANCOVA extension. How would you diagnose these problems?

END OF QUESTION (2)

3. A radiologist was interested in patients' preferences to the timing and methods of being informed with their diagnosis results. He considered the so-called *discrete choice*, a widely used method in marketing research. In particular, he designed an experiment with three categorical variables, each with three levels:

- Days = ("1", "4", "14"), the number of days between an image being taken and the diagnosis result being informed by either radiologist or patient's doctor
- When = (before, same, after), patient receives the diagnosis before, or the same day as, or after patient's doctor sees the diagnosis result summarized by radiologist
- Where = (portal, office, phone), patient receives the diagnosis result from portal (a secured website where the patient's medical information is stored), or in doctor's office, or from doctor's phone call

This particular discrete choice experiment consists of 12 choice sets, each set with two choices, thus in total there are 24 unique combinations of the levels of the above three categorical variables. A few sets are given below to help you understand the design:

Set	Choice	Day	When	Where
1	1	1	after	portal
1	2	3	same	office
2	1	1	same	phone
2	2	3	before	portal
⋮	⋮	⋮	⋮	⋮
12	1	14	same	phone
12	2	3	after	office

In the survey conducted by the radiologist, each patient selects one choice in each of the 12 sets. Denote C_{ijk} ($i = 1, \ldots, n$, $j = 1, \ldots, 12$, $k = 1, 2$) to be the choice made by patient i in set j for choice k, so $C_{ij1} = 1$ and $C_{ij2} = 0$ if the patient chooses Choice 1 in set j. The statistical model to analyze the above data is given below:

$$Pr(C_{ijk} = 1) = \frac{\exp\{\beta_1 \text{Days}_{jk} + \beta_2 \text{When}_{jk} + \beta_3 \text{Where}_{jk}\}}{\sum_{k=1}^{2} \exp\{\beta_1 \text{Days}_{jk} + \beta_2 \text{When}_{jk} + \beta_3 \text{Where}_{jk}\}}, \tag{1}$$

where each covariate is a categorical variable, and each of β_1, β_2, and β_3 is a vector of two regression coefficients.

(a) Instead of model (1), the radiologist ran a standard logistic regression. Explain what is wrong with such an analysis, and provide the correct method to analyze such data. You may consider using mock-up R code to help you describe your method.

(b) Based on model (1), interpret the estimating results given below:

Parameter	Estimate	Standard Error	P-Value
Day 1	1.7919	0.0797	< 0.0001
Day 3	1.4071	0.0753	< 0.0001
Day 14	0	.	.
After	0.0092	0.0714	0.8972
Before	-0.6244	0.0762	< 0.0001
Same	0	.	.
Office	0.2500	0.0695	0.0003
Phone	0.7117	0.0802	< 0.0001
Portal	0	.	.

(c) The radiologist also collected patients' demographic information such as age, gender, education, etc., and was interested in knowing how these demographic variables affect patients' preference for each of the three design variables. Provide details about how you would estimate the effects of the demographic variables.

(d) A common discrete choice design involves more than two choices in each choice set, for example, $k = 3$ for all the choice sets. What analysis method do you want to consider for such data? Justify your answer.

END OF QUESTION (3)

4. The *Young Citizens study* (Kamo et al., *American Journal of Public Health*, 2008) is a cluster or group randomized clinical trial (GRT) involving a behavioral intervention designed to train children ages $10-14$ years to educate their communities about HIV.

   A GRT assigns treatments to groups of individuals and is advantageous when interaction among subjects within a group (e.g. all subjects in a neighborhood) may impact their respective outcome. For example, in the *Young Citizens study*, **30** communities were grouped into **15** pairs based on some underlying characteristics. *One community per pair was randomly assigned to treatment and the other to control.* The underlying idea is that observations within a community are correlated.

   Residents within each community were surveyed post-intervention regarding their beliefs about the ability of children to effectively teach their peers and families about HIV. The primary outcome was a continuous composite score reflecting the strength of this belief (Y_1) and the goal of the analysis was to identify differences in training between treatment and control. The number of residents surveyed per community ranged from $16-80$ by multiples of 16. Here, we loosely follow the analysis in Stephens et al., *Statistics in Medicine*, $2012, 31, 915-930$.

   1. A standard approach for analyzing these data involves a linear mixed effects model of the form
   $$Y_{ij} = \beta_0 + \beta_1 A_i + b_i + \varepsilon_{ij},$$
   where Y_{ij} denotes the outcome for the j-th individual in the i-th cluster, A_i is an indicator for treatment (with $A_i = 1$ indicating the experimental treatment, and $A_i = 0$ the control), b_i is a random effect, and ε_{ij} denotes the measurement error, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$:

      a) Carefully detail all the assumptions characterizing the model above
      b) Provide *precise* interpretations of each parameter in the model and identify which parameter is of primary interest given the goals of the study.
      c) Characterize the asymptotic distribution of the maximum likelihood estimator $\hat{\beta}_{\text{MLE}} = (\hat{\beta}_0, \hat{\beta}_1)$.
      d) Propose a test for assessing the efficacy of the behavioral intervention. Clearly specify the hypotheses being tested, the hypothesis testing approach and the relevant test statistic.
      e) Propose a test for assessing the necessity of a random effect in the model. Clearly specify the hypotheses being tested, the hypothesis testing approach and the relevant test statistic.

   2. Another standard approach is based on the definition of a marginal model where the parameters' estimates are obtained as the solution of a system of first-order generalized estimating equations (GEE1).

f) Write down the first-order generalized estimating equation for the marginal model in this case. Each component of your estimating equation should be fully defined.

g) Stephens et al (2012) note that for cluster randomized designs, an independence covariance structure is generally assumed. Would you worry about the choice of an independence working correlation structure? Justify your answer, with particular regard to the impacts of this choice.

h) At some point in their manuscript, Stephens et al (2012) comment that the "sandwich variance estimator underestimates the variability of parameter estimates, and consequently results in inference that is too liberal". What might cause the robust variance estimator to perform poorly in this case? Justify your answer, taking into account the particular characteristics of the *Young Citizens study*.

i) Stephens et al (2012) propose an *augmented GEE* approach, whereby the usual set of estimating equations is modified by taking into account the probability of randomizing each community to either treatment or control, also as a function of known baseline covariates. More in detail, let $\pi_k = P(A = k|X)$ indicate the probability of being assigned to treatment $k = 1, \ldots, K$. Here $K = 2$. Then, Stephens et al (2012) propose the general form of the augmented GEE as

$$\psi\left(\mathbf{Y}, \mathbf{X}; \beta\right) - \sum_{i=1}^{m} \sum_{k=1}^{K} \left\{ I\left(A_i = k\right) - \pi_k \right\} \gamma_k\left(\mathbf{X}_i\right) = \mathbf{0}$$

where $\psi\left(\mathbf{Y}, \mathbf{X}; \beta\right)$ denotes the estimating function already considered in the first-order GEE1 at point (f) above, and $\gamma_k\left(\mathbf{X}_i\right)$ is a p -dimensional function of $\mathbf{X}_i$. Assume $\pi_k = P(A = k|X)$ to be known. Under regularity conditions, the estimator obtained as the solution of the augmented GEE is consistent and asymptotically normal. Based on the general theory of GEEs, provide a crucial condition for such result to hold.

3. A secondary outcome measured residents' beliefs regarding whether or not the AIDS problem was getting worse in their communities (Y_2). Responses were dichotomized into one category "agree" (1) and one category "disagree" (0).

j) Consider a marginal model such that $E\left(Y_{2,ij}|A_i\right) = g\left(A_i; \beta\right) = g\left(\beta_0 + \beta_1 A_i\right)$. Clearly outline the specification of the mean and variance components of the model.

k) The following table contains the marginal treatment effect analysis that was conducted with the binary outcome Y_2.

Estimator	$\hat{\beta}_1$	SE	95%CI
Exch	-0.238	0.275	$(-0.777, 0.300)$

Provide an interpretation of the inference for the coefficient β_1 in the evaluation of the association between treatment and the outcome Y_2.

END OF QUESTION (4)